1. What are the key tasks that machine learning entails? What does data pre-processing imply?

Machine Learning involves several key tasks which can be categorized as follows:

* Data collection
* Data pre-processing
* Feature Engineering
* Model Selection and training
* Hperparameter tunning
* Deployment and monitoring

Data Pre-processing is a crucial step in machine learning. It involves transforming raw data into a format that is suitable for training machine learning models. The main goals of data pre-processing are to ensure data quality, remove inconsistencies, and make the data compatible with the chosen machine learning algorithm.

2. Describe quantitative and qualitative data in depth. Make a distinction between the two.

Quantitative and qualitative data are two distinct types of data that are commonly used in research and analysis. Here's a detailed explanation of each:

i) Quantitative Data:

- Definition: Quantitative data refers to information that is expressed in numerical form and can be measured or counted. It involves data that can be quantified and analyzed using mathematical and statistical methods.

- Characteristics:

- Numeric values: Quantitative data consists of numbers that represent quantities or measurements.

- Continuous or discrete: Quantitative data can be continuous, meaning it can take any value within a specific range (e.g., temperature, time), or discrete, where values are distinct and separate (e.g., number of products sold).

- Objective and standardized: Quantitative data is typically objective and can be measured consistently using standardized methods or instruments.

- Statistical analysis: Quantitative data lends itself well to statistical analysis, allowing researchers to identify patterns, relationships, and trends.

- Examples: Height, weight, temperature, income, test scores, number of customers.

ii) Qualitative Data:

- Definition: Qualitative data refers to non-numerical information that is descriptive in nature and typically captures subjective characteristics, opinions, behaviors, or qualities.

- Characteristics:

- Descriptive and subjective: Qualitative data provides descriptive information about qualities, behaviors, experiences, and opinions, often in the form of words, texts, or images.

- Non-standardized: Qualitative data collection methods are typically flexible and context-dependent, allowing for a deeper exploration of specific phenomena.

- Richness and depth: Qualitative data often captures rich and detailed information, providing insights into the complexities of human behavior and experiences.

- Interpretive analysis: Qualitative data requires interpretive analysis techniques, such as coding, thematic analysis, or narrative analysis, to identify patterns, themes, and meanings.

- Examples: Interview transcripts, open-ended survey responses, observations, focus group discussions, case studies, textual data.

Distinction between Quantitative and Qualitative Data:

i). Nature of data: Quantitative data consists of numeric values and can be measured, counted, and analyzed using mathematical and statistical methods. Qualitative data, on the other hand, is non-numerical and focuses on descriptive information, subjective experiences, and contextual insights.

ii) Measurement: Quantitative data involves objective and standardized measurements, while qualitative data involves subjective interpretation and non-standardized data collection methods.

3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

i) Quantitative Data:

- Attribute: Age (numeric)

- Attribute: Income (numeric)

- Attribute: Number of products purchased (numeric)

Example:

- Age: 32

- Income: $50,000

- Number of products purchased: 5

ii) Qualitative Data:

- Attribute: Gender (categorical: Male/Female/Other)

- Attribute: Marital status (categorical: Single/Married/Divorced)

- Attribute: Education level (categorical: High School/Diploma/Bachelor's/Master's)

Example:

- Gender: Female

- Marital status: Married

- Education level: Bachelor's

iii) Time Series Data:

- Attribute: Timestamp (date and time)

- Attribute: Stock price (numeric)

Example:

- Timestamp: 2023-05-23 14:30:00

- Stock price: $95.50

iv). Textual Data:

- Attribute: Customer reviews (text)

- Attribute: Product descriptions (text)

Example:

- Customer reviews: "Great product! Highly recommended."

- Product descriptions: "This laptop features a powerful processor and a sleek design."

The data collection record can be as follows:

Record 1:

- Age: 32

- Income: $50,000

- Number of products purchased: 5

- Gender: Female

- Marital status: Married

- Education level: Bachelor's

- Timestamp: 2023-05-23 10:15:00

- Stock price: $100.20

- Customer reviews: "Great product! Highly recommended."

- Product descriptions: "This laptop features a powerful processor and a sleek design."

4. What are the various causes of machine learning data issues? What are the ramifications?

Machine learning data can be affected by various issues that can impact the performance and reliability of machine learning models. Here are some common causes of machine learning data issues:

* Insufficient or biased data
* Inconsistent or noisy data
* Missing data
* Null type data
* Imbalanced Data
* Data Leakage

To mitigate these issues, it is crucial to prioritize data quality, perform thorough data pre-processing, handle missing data appropriately, address data imbalance, and regularly monitor and update models to account for data drift. Furthermore, careful consideration should be given to the fairness and ethical implications of the data used for training machine learning models.

5. Demonstrate various approaches to categorical data exploration with appropriate examples.

i) Equating the data with object type data-type. Ex: data[‘feature’]!=’O’

ii) Plotting different plots like bar plot and pie plots

iii) Association measures, such as chi-square test or Cramér's V, help assess the strength and significance of relationships between categorical variables.

6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

Missing value would be disastrous in learning activity. The machine may not be able to train itself if missing values are present. Absurd and ambiguous output may be found if such kind of dataset is used. Hence it is always advised to check for missing values while approaching ML algorithm.

Missing values need to be checked before rushing to any ML algorithm. These can be solved by :

i) Imputation: The missing values can be imputed by mean/median/mode values.

ii) Deletion: Delete the missing values

iii) Analyse the missing values and impute with proper values

7. Describe the various methods for dealing with missing data values in depth.

Handling missing values is an important step in machine learning (ML) data preprocessing. Here are several approaches you can use to handle missing values:

* Identify missing values: Begin by identifying the missing values in your dataset. They may be represented by NaN (Not a Number), NULL, NA, or other placeholders.
* Analyze the reasons for missingness: Understand why the values are missing. Missing data can occur due to various reasons, such as data entry errors, equipment malfunction, or deliberate omissions. Knowing the reasons can help determine the appropriate handling strategy.
* Deletion: If the missing values are relatively few, you may choose to delete the corresponding rows or columns. This approach is suitable when missing data does not introduce significant bias or information loss. However, it should be used with caution to avoid removing too much valuable data.
* Imputation: Imputation involves filling in the missing values with estimated or calculated values. Several imputation techniques are commonly used:
* Mean/Median/Mode imputation: Replace missing values with the mean, median, or mode of the available values in the same feature. This method assumes that the missing values are missing completely at random (MCAR) or missing at random (MAR).

* + Regression imputation: Predict missing values using regression models based on other features. For each missing value, a regression model is built using the other features as predictors, and the missing value is then imputed based on the model's prediction.

* + Hot-deck imputation: Replace missing values with randomly selected values from similar records or observations in the same dataset.

* + Multiple imputatio: Create multiple imputations by using statistical models to estimate missing values multiple times. This method accounts for uncertainty in imputation and can improve the accuracy of subsequent analyses.

* + Domain-specific imputation: Utilize domain knowledge or business rules to impute missing values. For example, if missing values are related to a specific category, you may assign a specific value based on the category.
* Flagging: Create an additional binary column indicating whether a value is missing or not. This approach can help the ML algorithm learn patterns associated with missing values.
* Advanced techniques\*\*: Advanced techniques such as k-nearest neighbors (KNN) imputation, decision trees, or random forests can be employed to impute missing values based on patterns in the data.

9

i. What is the IQR? What criteria are used to assess it?

IQR is Inter Quartile Range. The IQR (Interquartile Range) is a statistical measure that represents the range between the first quartile (Q1) and the third quartile (Q3) of a dataset. It is a robust measure of dispersion that is less sensitive to outliers compared to the range or standard deviation.

Criteria used to acess are:

i) Detectig outliers

ii) plotting box plot

iii) Skewness and symmetry

iv) Data Variability

ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?

A box plot, also known as a box-and-whisker plot, is a graphical representation that displays the distribution of a dataset through several components. Each component of a box plot provides information about different statistical measures and characteristics of the data. The main components of a box plot include:

* **Minimum and maximum values**: The whiskers of the box plot represent the minimum and maximum values within a certain range. By default, they extend up to 1.5 times the IQR (Interquartile Range) from the first quartile (Q1) and the third quartile (Q3), respectively.
* **Lower whisker**: The lower whisker extends from the box to the minimum value within the defined range.
* **Upper whisker**: The upper whisker extends from the box to the maximum value within the defined range.
* **Median**: The median, also known as the second quartile (Q2), is represented by a line inside the box. It divides the dataset into two equal halves, indicating the central tendency of the data**.**
* **Interquartile Range (IQR):** The IQR is the range between Q1 and Q3, representing the middle 50% of the dataset. It is represented by the width of the box.

When the lower whisker surpasses the upper whisker in length, it means that the dataset is highly skewed or asymmetric. This situation occurs when the values in the lower half of the dataset (between Q1 and the minimum value) are more spread out compared to the upper half (between Q3 and the maximum value).

**Box plots can be used to identify outliers in the following ways:**

* Outlier definition**:** By default, box plots consider data points beyond 1.5 times the IQR as potential outliers. These points are plotted individually outside the whiskers, allowing easy identification of potential extreme values.
* Visual identification: Box plots provide a visual representation of the dataset's distribution. Points outside the whiskers are easily identifiable and can be visually examined for potential outliers.
* Comparison to whisker lengths: If the whiskers of the box plot are of unequal lengths, with one whisker extending significantly further than the other, it suggests the presence of outliers in the corresponding direction.
* Custom outlier definition: Box plots can be customized to use different criteria for identifying outliers. By adjusting the range beyond the whiskers or using other statistical methods, specific definitions of outliers can be applied to suit the data and analysis requirements.

10. Make brief notes on any two of the following:

1. Data collected at regular intervals

Data collected at regular intervals refers to a type of data collection where observations or measurements are recorded at consistent and uniform time intervals. This regular interval can be daily, weekly, monthly, yearly, or any other fixed duration.

2. The gap between the quartiles

A gap between quartiles refers to a significant difference or distance between the values of the first quartile (Q1) and the third quartile (Q3) in a dataset. The quartiles divide the data into four equal parts, with Q1 representing the 25th percentile and Q3 representing the 75th percentile. Analyzing the gap between quartiles provides insights into the distribution and spread of the data. A larger gap indicates greater variability or skewness, while a smaller gap suggests a more concentrated or uniform distribution. It is essential to consider the reasons behind the gap and investigate further to understand the underlying patterns and characteristics of the data.

3. Use a cross-tab

1. Make a comparison between:

1. Data with nominal and ordinal values

**Nominal values**: Nominal values represent categories or labels that do not have a specific order or numerical value associated with them. Examples of nominal variables include gender (male/female), marital status (single/married/divorced), or categorical variables like color (red/blue/green).

**Ordinal values:** Ordinal values represent categories or labels that have a specific order or rank associated with them. Examples of ordinal variables include ratings (e.g., Likert scales), education level (e.g., high school, bachelor's, master's), or satisfaction levels (e.g., very dissatisfied, dissatisfied, neutral, satisfied, very satisfied).

2. Histogram and box plot

**Histogram:** A histogram is a graphical representation of the distribution of continuous or discrete data. It consists of a series of adjacent rectangles (bars), where the area of each bar represents the frequency or proportion of data points falling within a particular interval or bin.

**Boxplot:** A box plot is a graphical representation that provides a summary of the distribution of continuous or ordinal data. It displays key statistical measures such as the minimum, maximum, quartiles, and potential outliers.

3. The average and median

**Average:** The term "average" refers to a broad concept of summarizing a set of values or observations. It is a general term used to describe various measures of central tendency that aim to represent the typical value or central value of a dataset.

**Mean**: The mean is a specific measure of central tendency and is one of the most commonly used averages. It is calculated by summing up all the values in a dataset and dividing the sum by the total number of observations. The mean represents the arithmetic average of the dataset and provides a measure of the "typical" value.